**

**HUNAN UNIVERSITY**

《数据仓库技术与应用》

实验报告八

|  |  |
| --- | --- |
| **报告名称：** | 基于信用卡逾期数据的Spark数据处理与分析 |
| **学生姓名：** | 杨超然 |
| **学生学号：** | 202106060220 |
| **专业班级：** | 电商2102班 |
| **学 院：** | 工商管理学院 |
| **指导老师：** | 周中定 |
| **日 期：** | 2023.4.10 |

目录

[一、实验过程 3](#_Toc739802400)

[（一）实验环境配置 3](#_Toc2076104317)

[（二） 数据集处理 3](#_Toc770959363)

[2.1数据集下载 3](#_Toc1745171590)

[2.2数据预处理 4](#_Toc767262404)

[2.3将文件上传至HDFS文件系统 5](#_Toc1887407440)

[（三） 使用Spark对数据处理分析 6](#_Toc1175894243)

[（四） 数据可视化 9](#_Toc2110022407)

[4.1可视化工具与代码 9](#_Toc1749131538)

[（五） 可视化结果展示 12](#_Toc750115383)

[5.1总体展示： 13](#_Toc1460234191)

[5.2各年龄段逾期人数展示： 13](#_Toc712930221)

[5.3逾期记录展示： 16](#_Toc1406957734)

[5.4房地产抵押数量展示 17](#_Toc796198221)

[5.5家属人数展示 19](#_Toc732895890)

[二、易错点总结 22](#_Toc1962507685)

[（一）无法执行pyecharts安装命令 22](#_Toc665327522)

[（二）pyecharts安装无响应 23](#_Toc212312325)

[（三）无法读取原始数据进行预处理 23](#_Toc1362908608)

[（四）处理后数据无法保存至本地 24](#_Toc1344395754)

[（五）无法将代码文件上传至spark-submit中 25](#_Toc1583987391)

[三、学习小记 25](#_Toc1868792325)

[（一）实验心得 25](#_Toc1844268900)

[（二）知识点笔记 26](#_Toc1995925149)

[2.1Pyecharts 26](#_Toc1819413103)

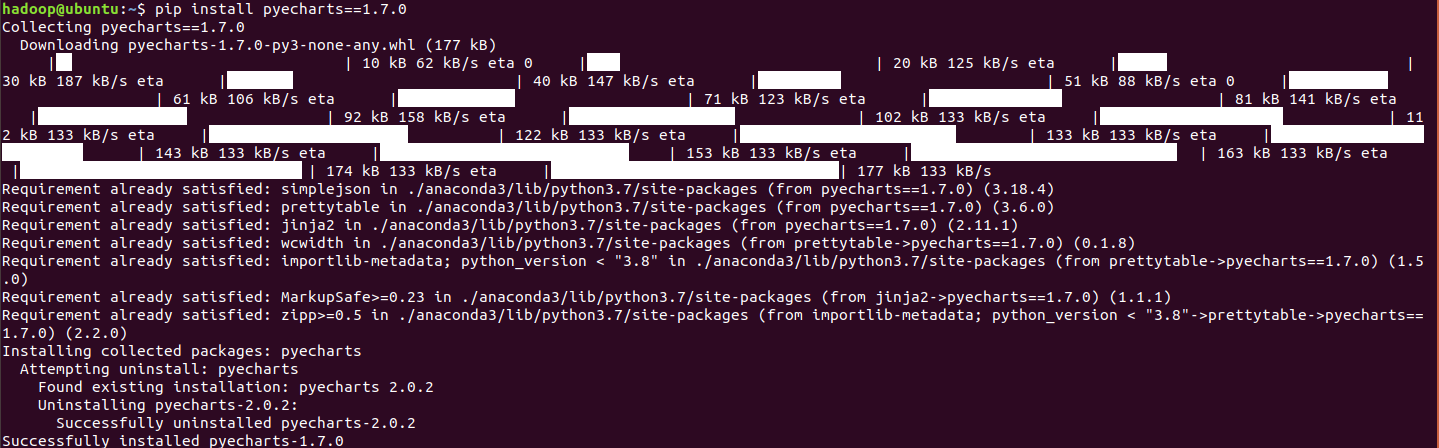
[2.2pip、pip3等命令超时的解决办法 27](#_Toc856372488)

# 一、实验过程

## （一）实验环境配置

本次实验环境：Linux：Ubuntu 16.04；Hadoop3.1.3；Python：3.6；Spark：2.4.0；Jupyter Notebook 相关安装操作见实验六。

除此之外，还需安装相关插件以支持python可视化分析。

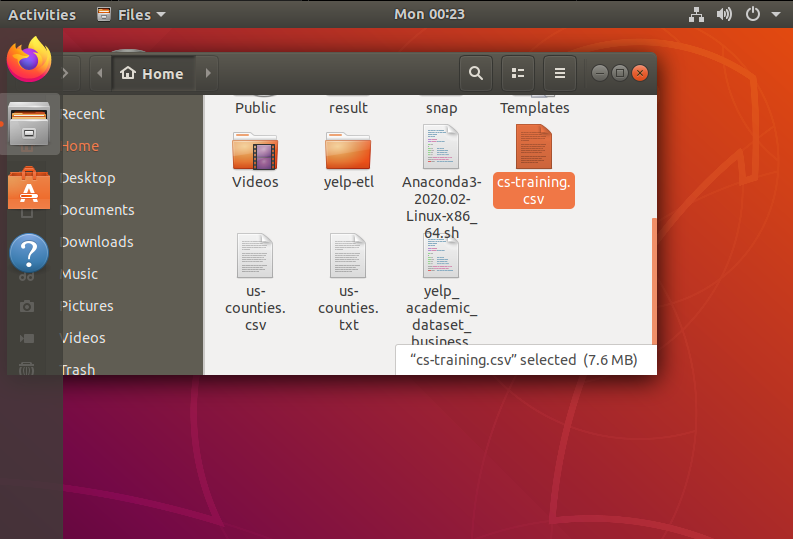


## 数据集处理

2.1数据集下载

本次实验数据集来自和鲸社区的信用卡评分模型构建数据，分析主体为数据集cs-training.csv，每条数据包含以下字段：

1. SeriousDlqin2yrs 是否逾期 0,1  
    （2）RevolvingUtilizationOfUnsecuredLines 信用卡和个人信贷额度的总余额 0.766126609  
    （3）Age 年龄 45,20,30  
    （4）NumberOfTime30-59DaysPastDueNotWorse 借款人逾期30-59天的次数 0,2,3  
    （5）DebtRatio 负债比率 0.802982129  
    （6）MonthlyIncome 月收入 9120,3000  
    （7）NumberOfOpenCreditLinesAndLoans 未偿还贷款数量 ,0,4,13  
    （8）NumberOfTimes90DaysLate 借款人逾期90天以上的次数 0,1,3  
    （9）NumberRealEstateLoansOrLines 房地产贷款的数量 3,6  
    （10）NumberOfTime60-89DaysPastDueNotWorse 借款人逾期60-89天的次数 0,3  
    （11）NumberOfDependents 家庭中的家属人数 0,1,3

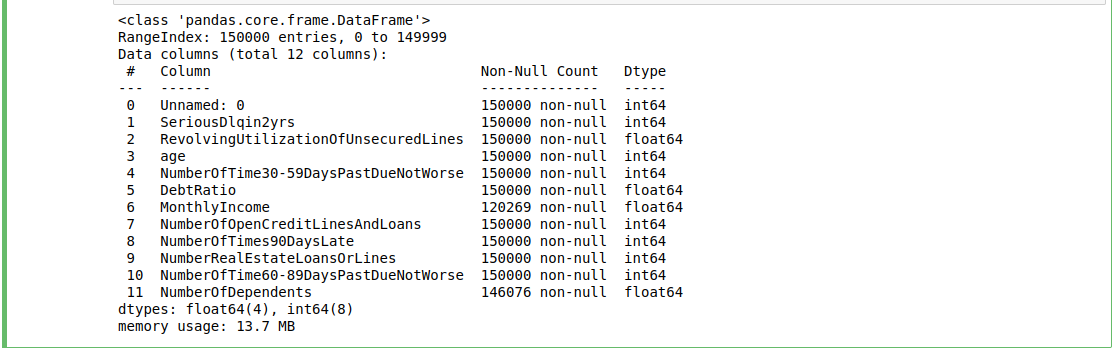


2.2数据预处理

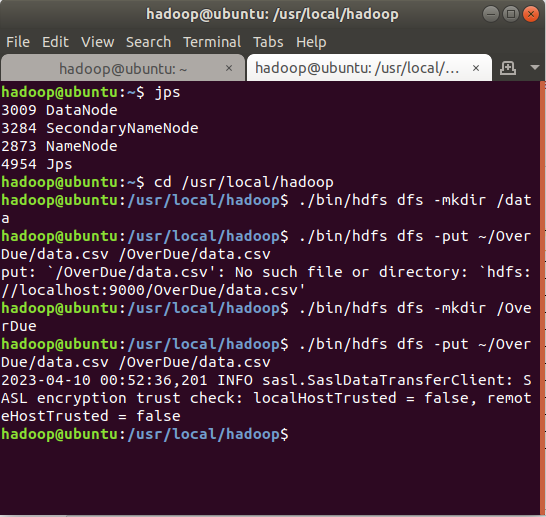
本次实验采用pandas库对数据进行预处理。在实验中，不对信用卡和个人信贷额度的总余额、负债比率、未偿还贷款数量、逾期90天以上的次数这4个属性进行处理分析。  
 具体处理步骤如下：  
 （1）读取数据  
 （2）查看数据是否具有重复值，去除重复值  
 （3）查看各字段缺失率，缺失值以均值填充  
 （4）选取要研究的属性，删除不研究的属性  
 （5）保存文件到本地



运行成功后结果如下：



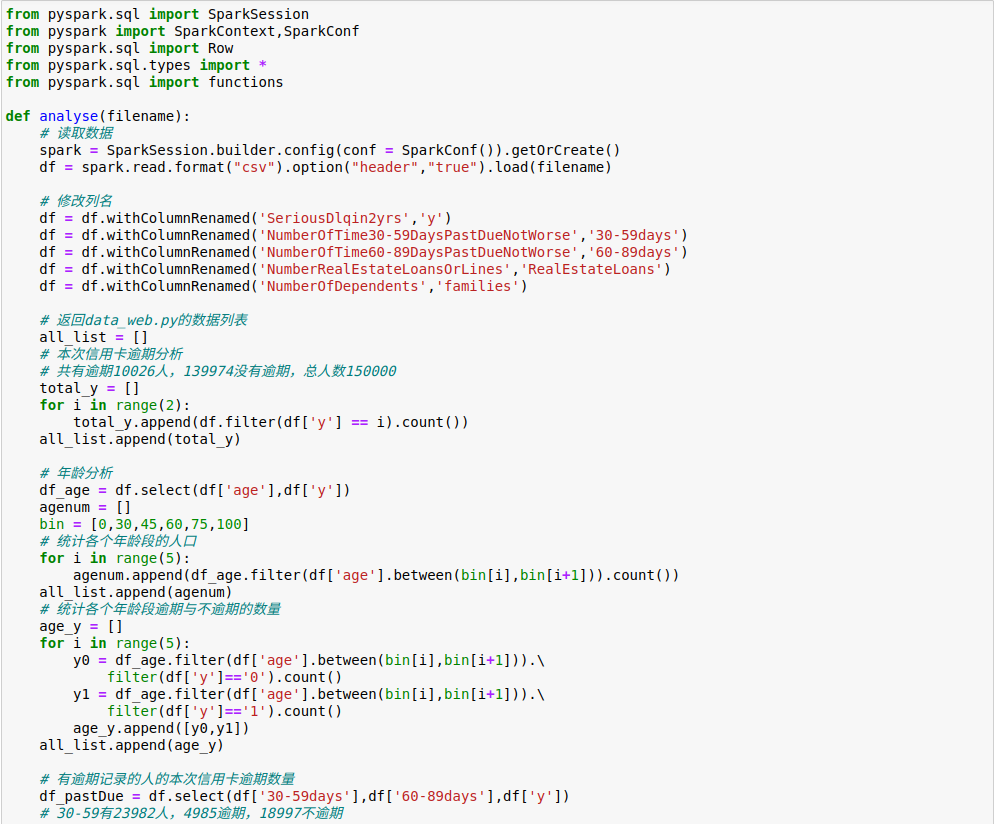
2.3将文件上传至HDFS文件系统

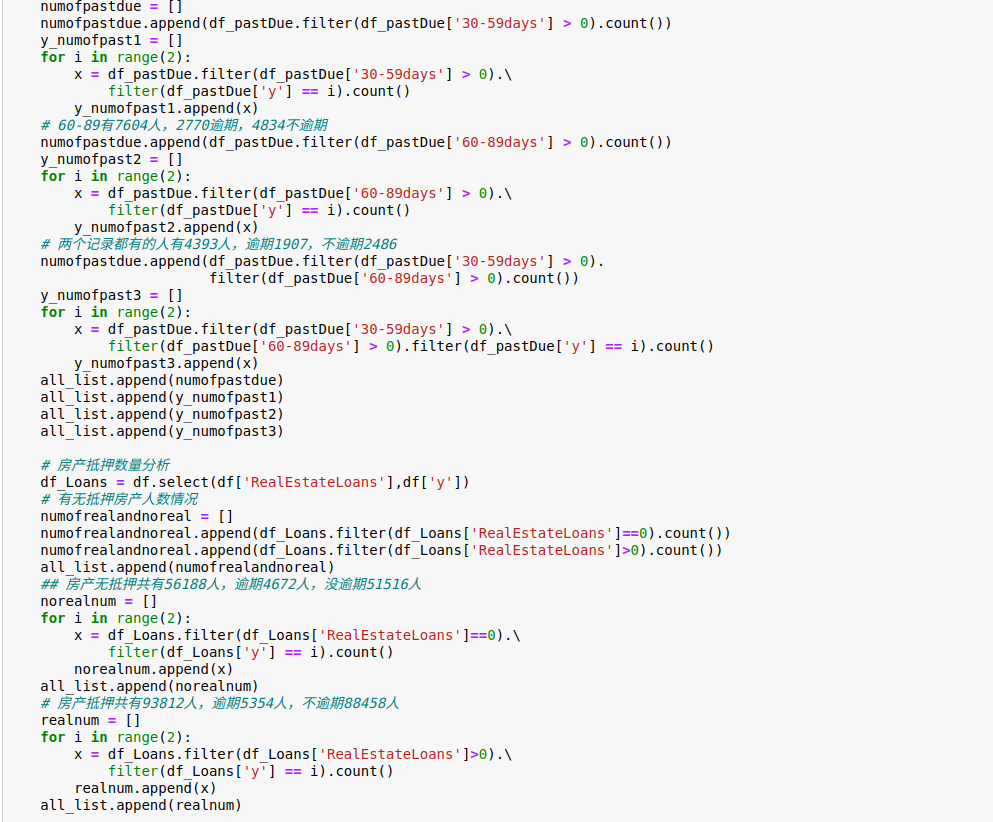


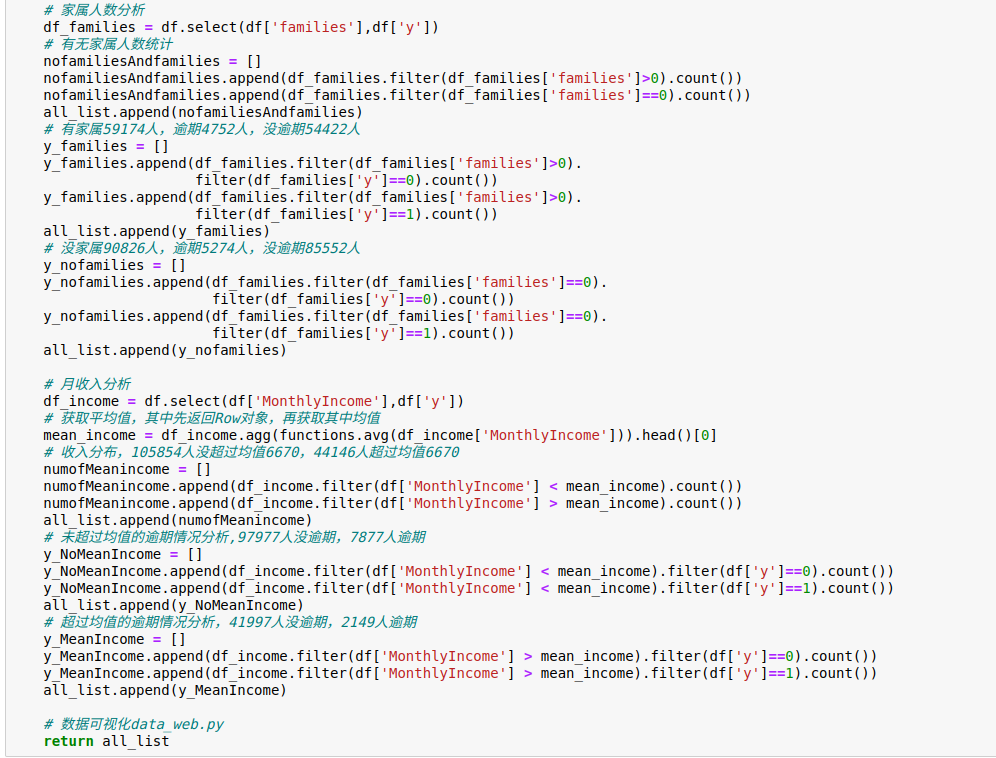
## 使用Spark对数据处理分析

我们将采用Python编程语言和Spark大数据框架对数据集“data.csv”进行处理分析，具体步骤如下：  
 （1）读取HDFS文件系统中的数据文件，生成DataFrame  
 （2）修改列名  
 （3）本次信用卡逾期的总体统计  
 （4）年龄与本次信用卡逾期的结合统计  
 （5）两次逾期记录与本次信用卡逾期的结合统计  
 （6）房产抵押数量与本次信用卡逾期的结合统计  
 （7）家属人数与本次信用卡逾期的结合统计  
 （8）月收入与本次信用卡逾期的结合统计  
 （9）将统计数据返回给数据可视化文件data\_web.py

代码如下：



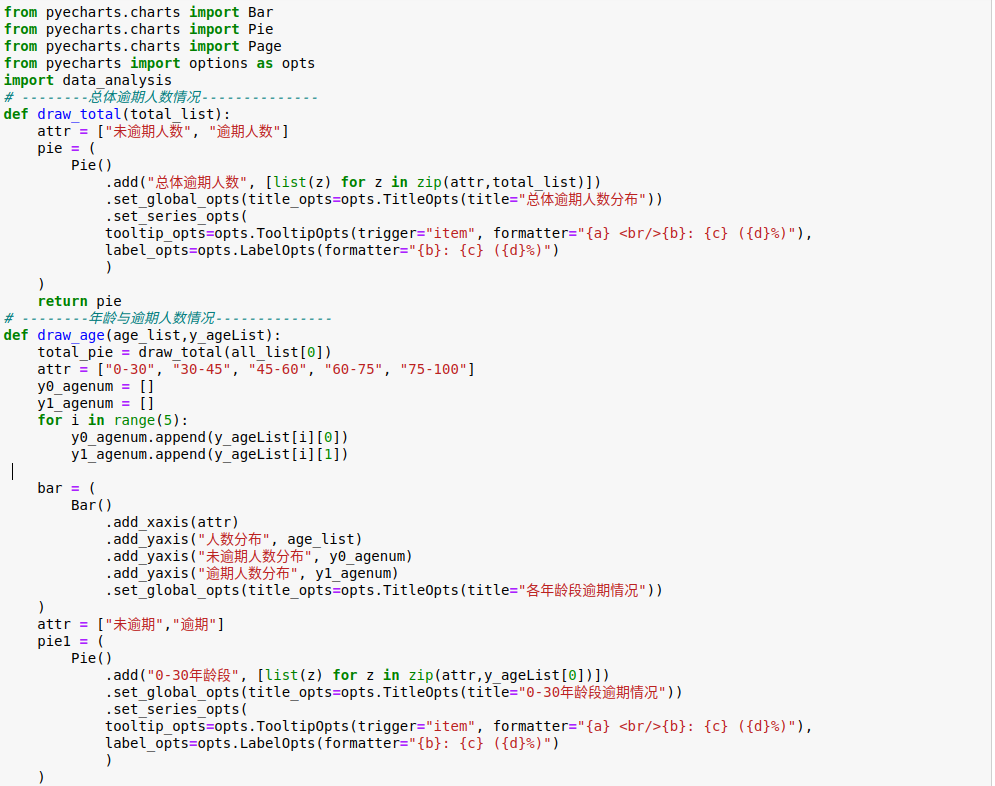




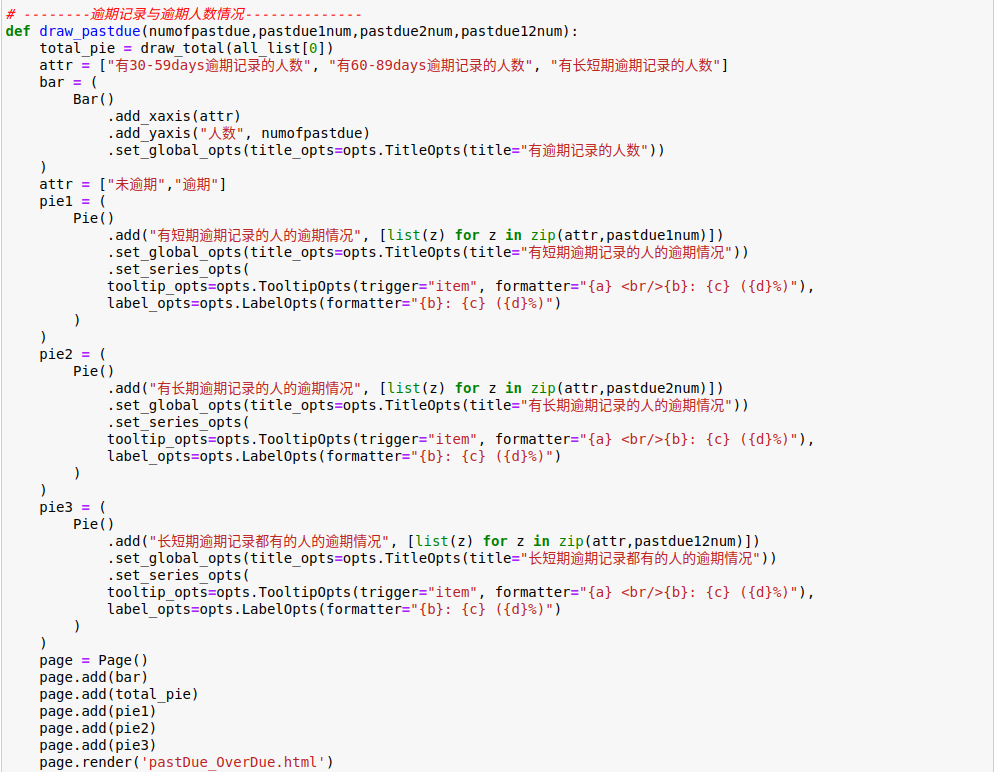
## 数据可视化

4.1可视化工具与代码

其中data analysis模块对应def(analysis)

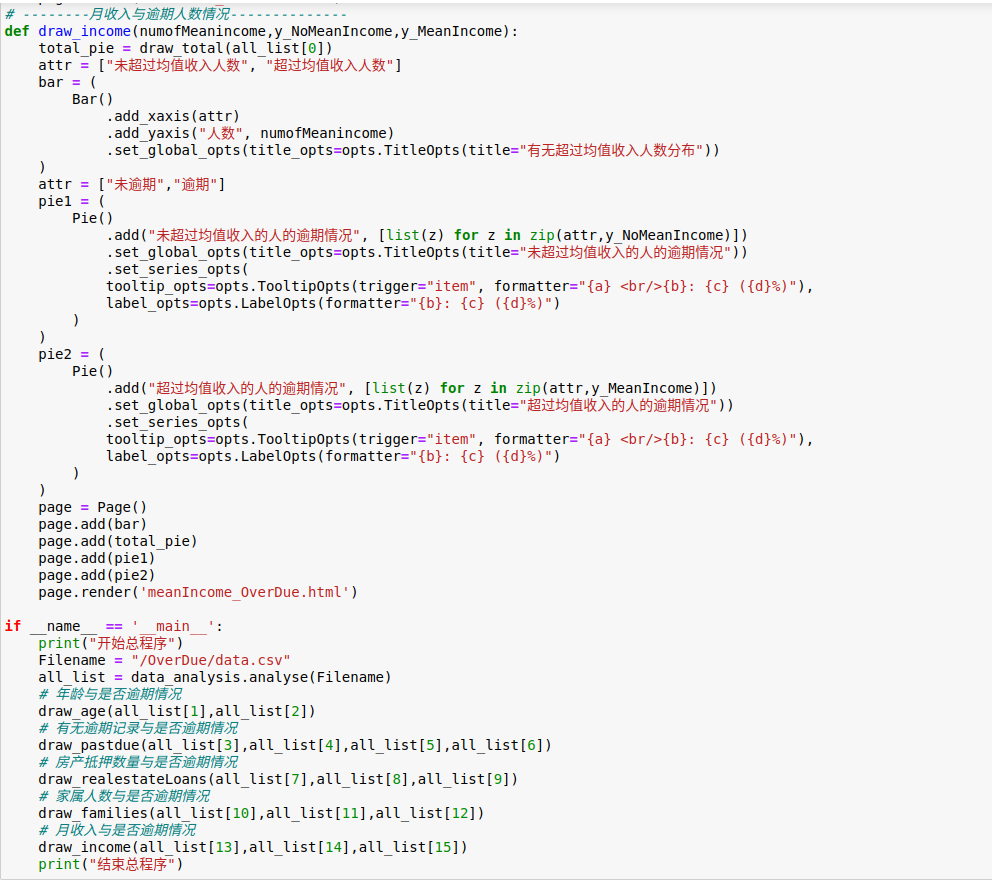










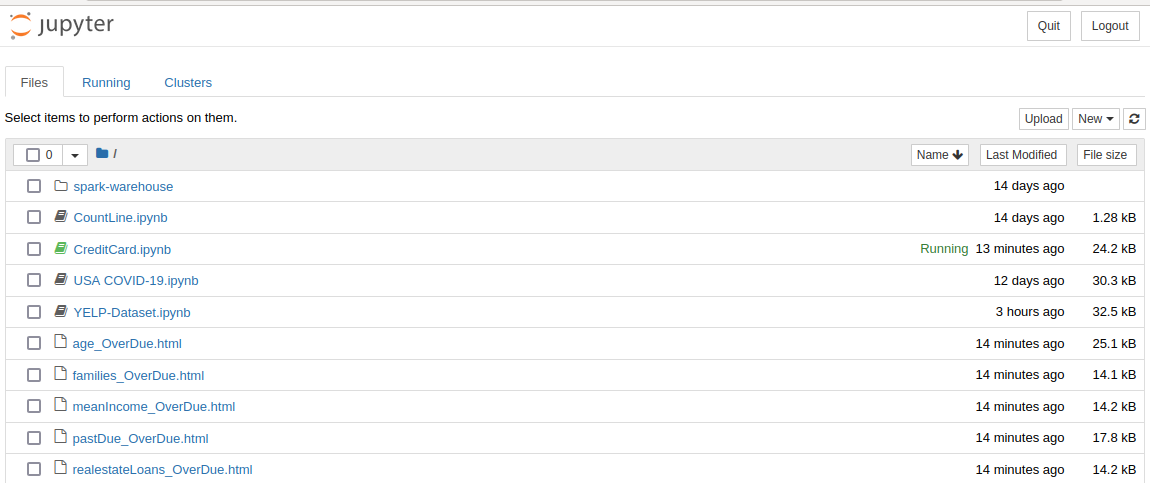


代码运行结果如下：

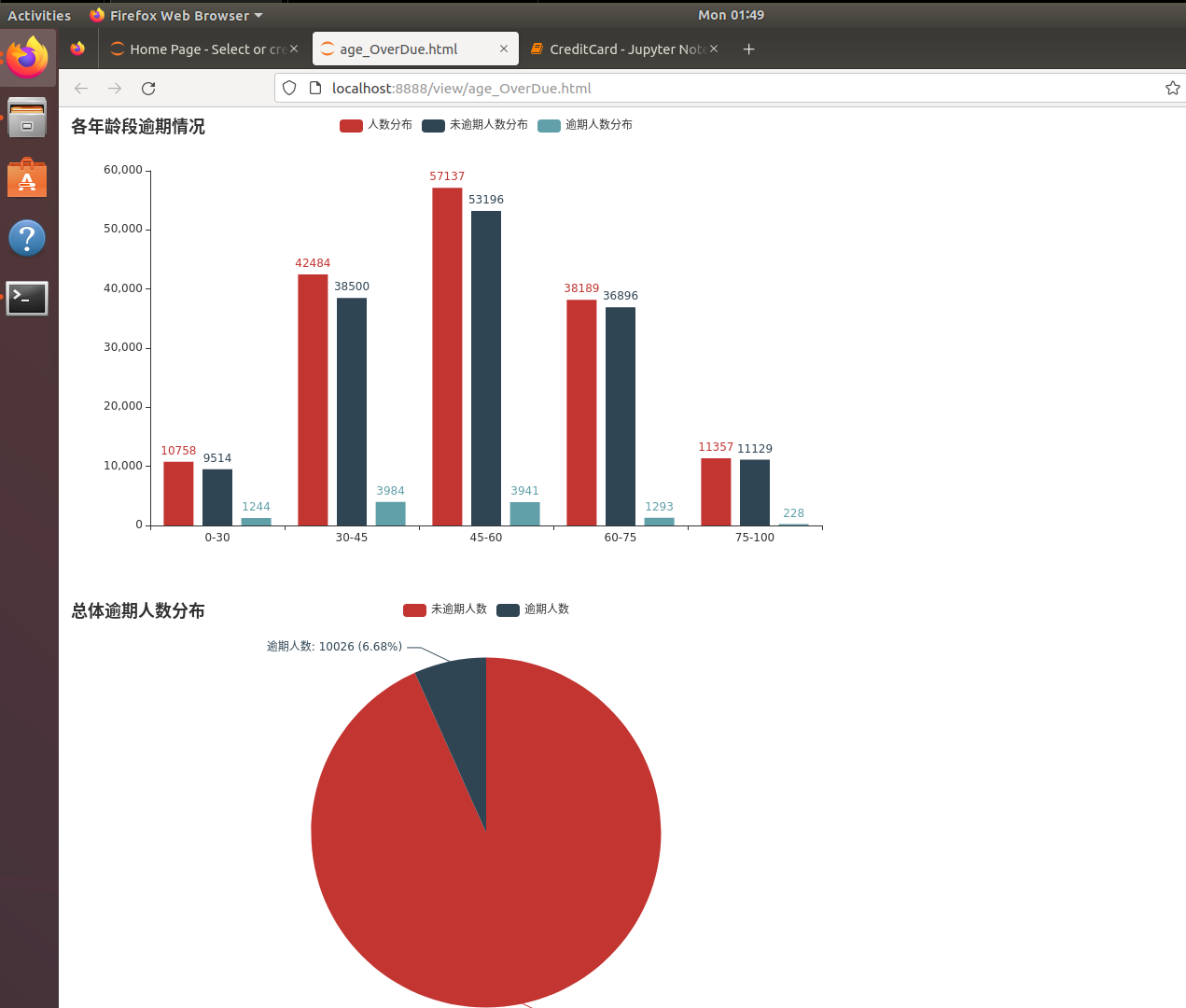


## 可视化结果展示

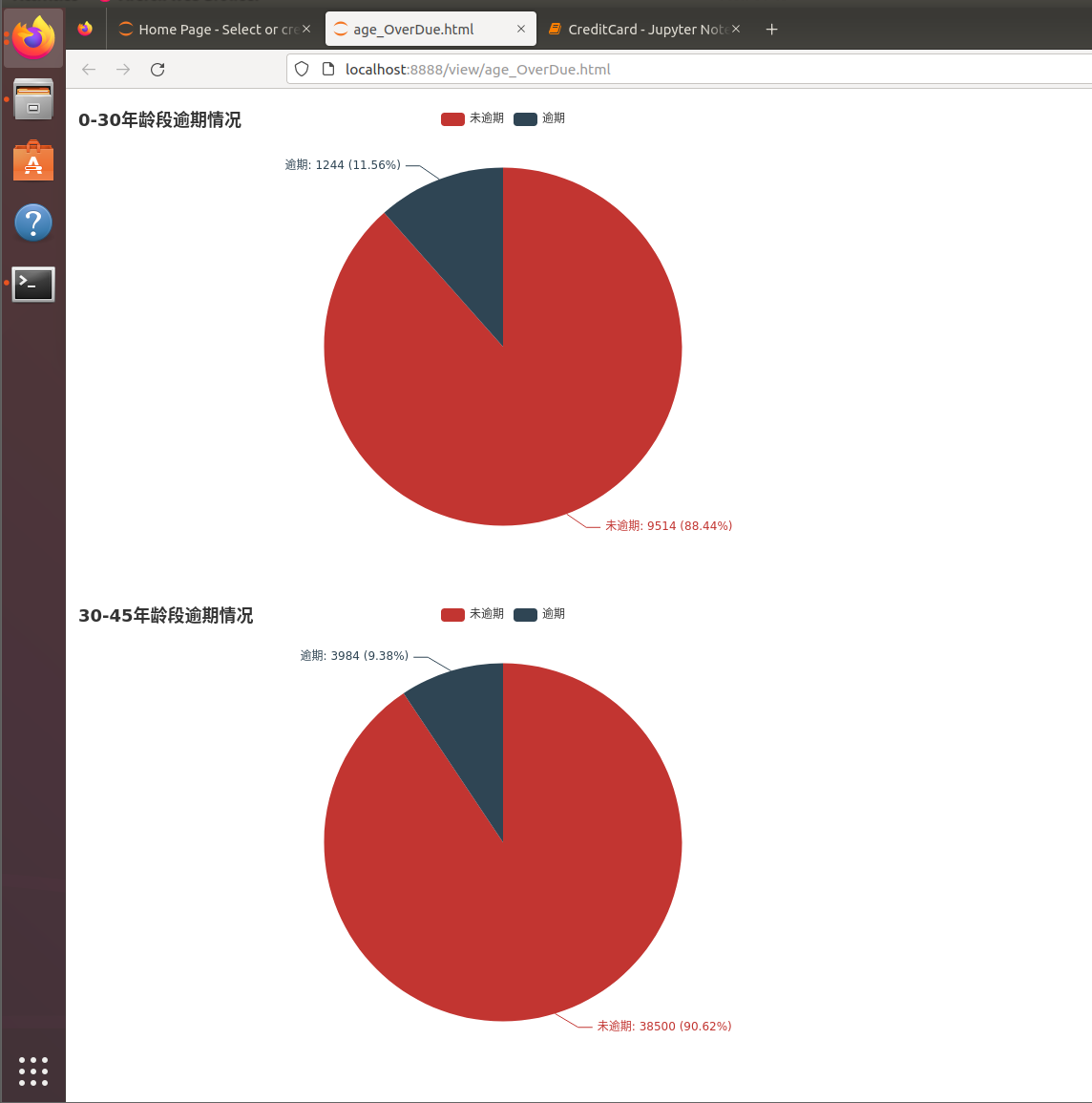
代码运行完毕后，Jupyter Notebook的Home Page中会出现5个html结果文件，即为我们的数据分析结果：

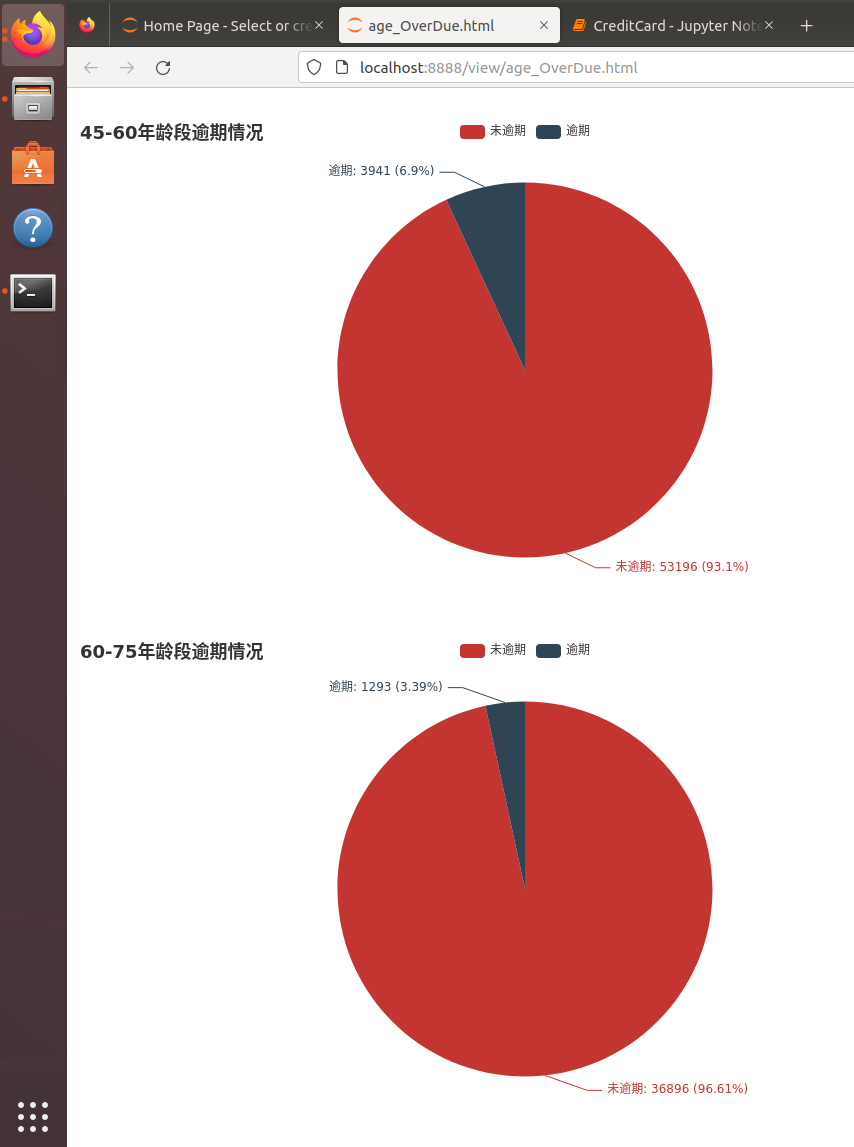


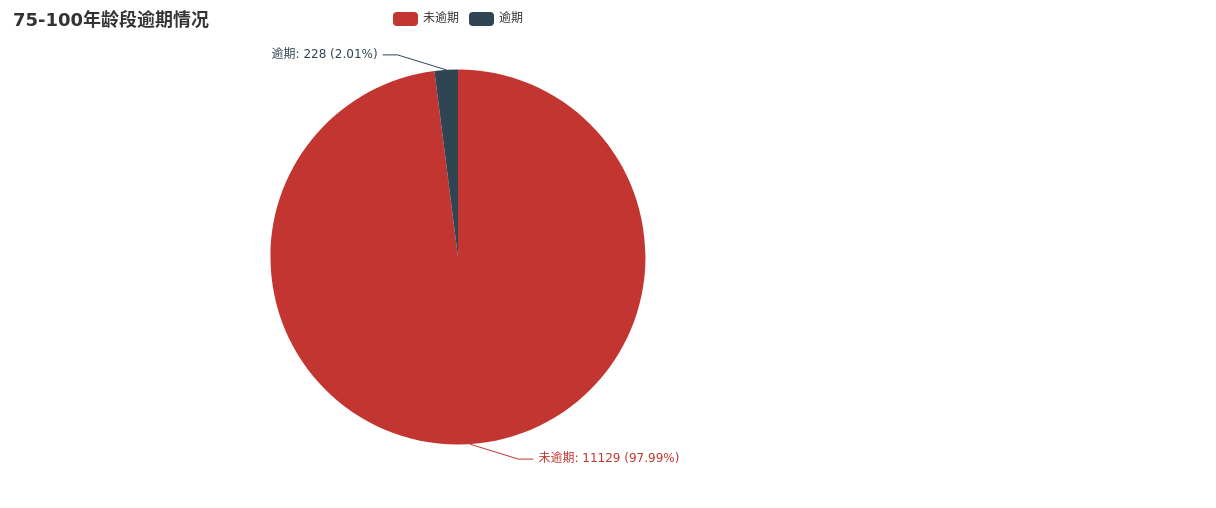
5.1总体展示：



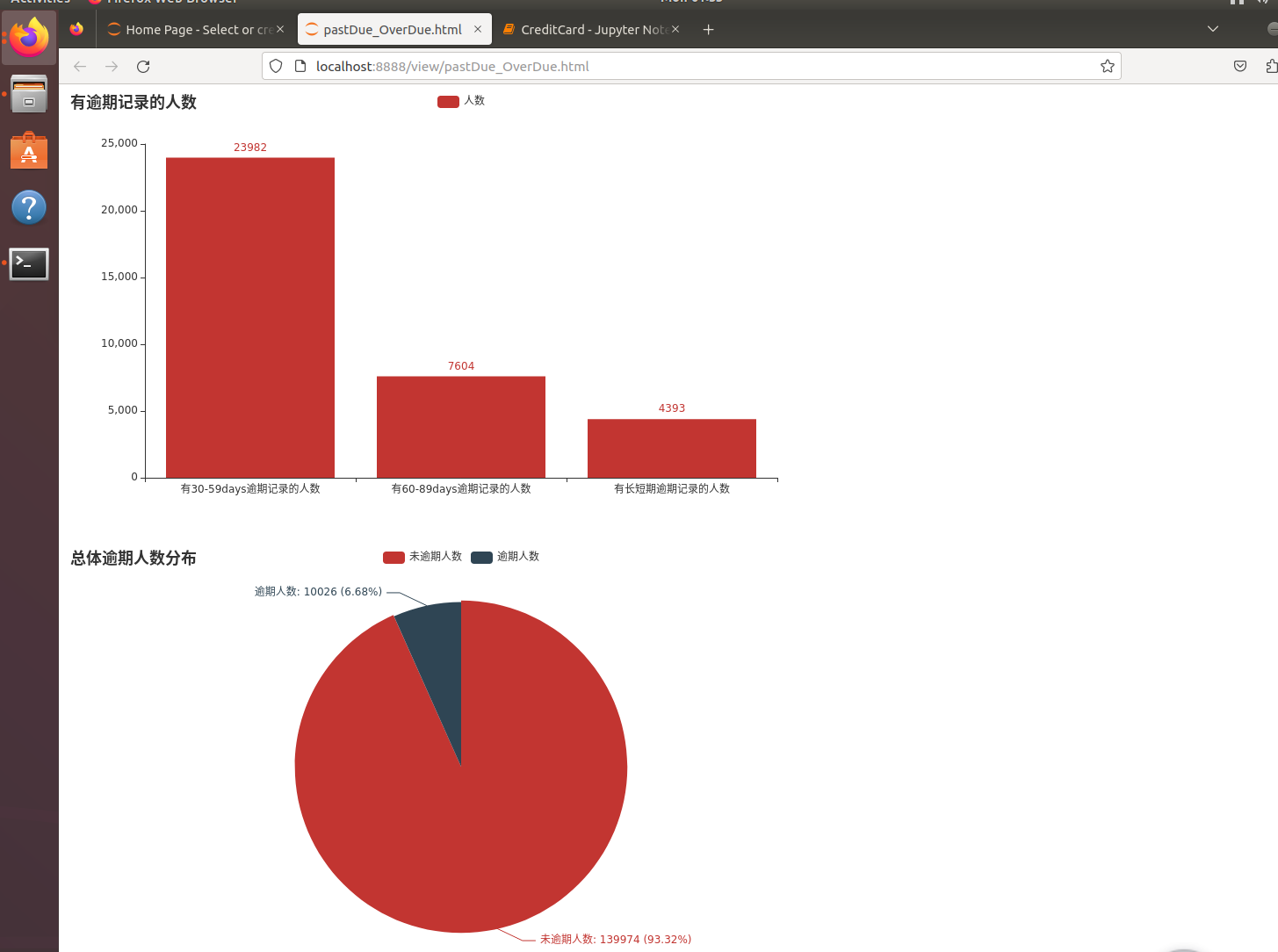
5.2各年龄段逾期人数展示：

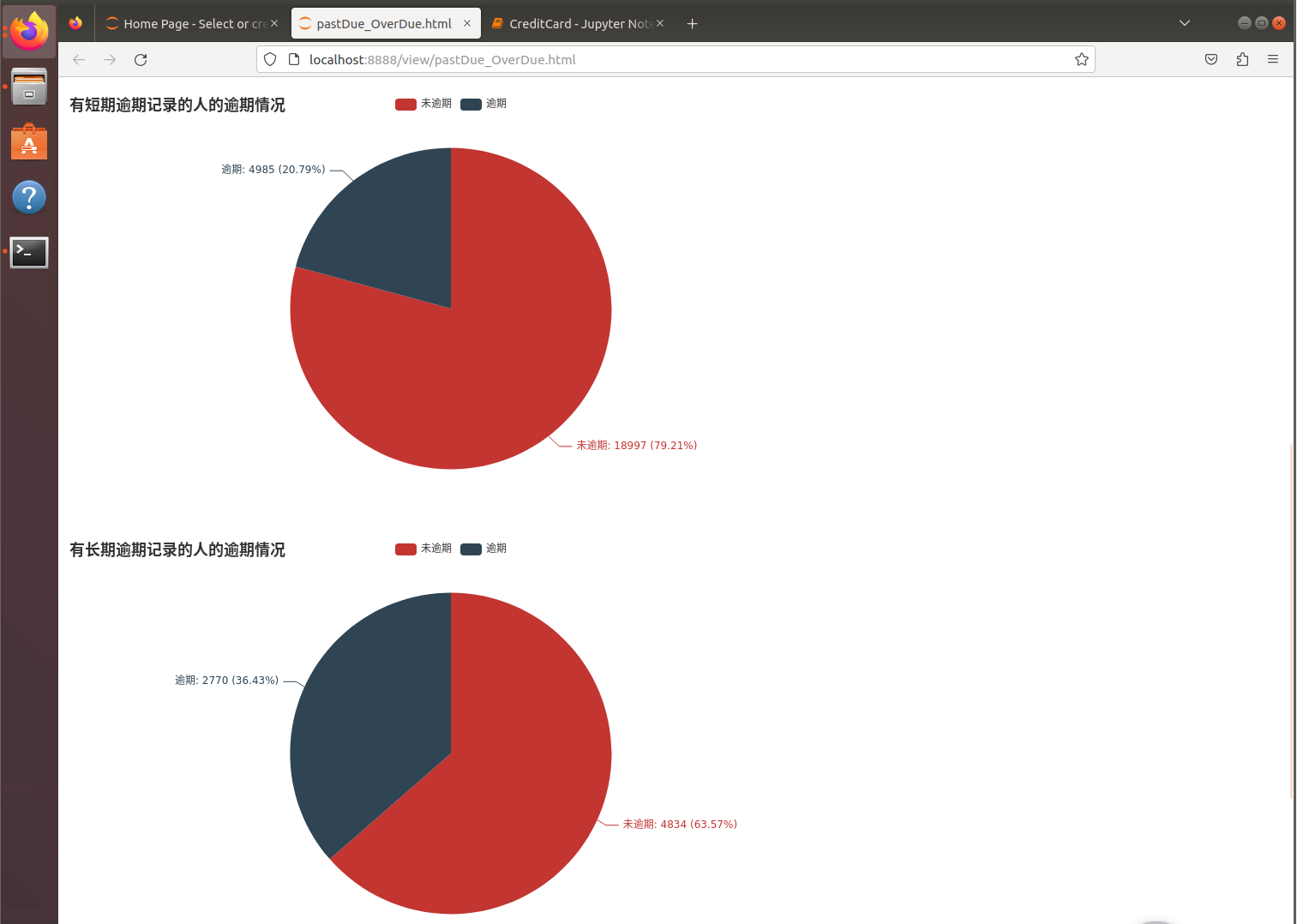


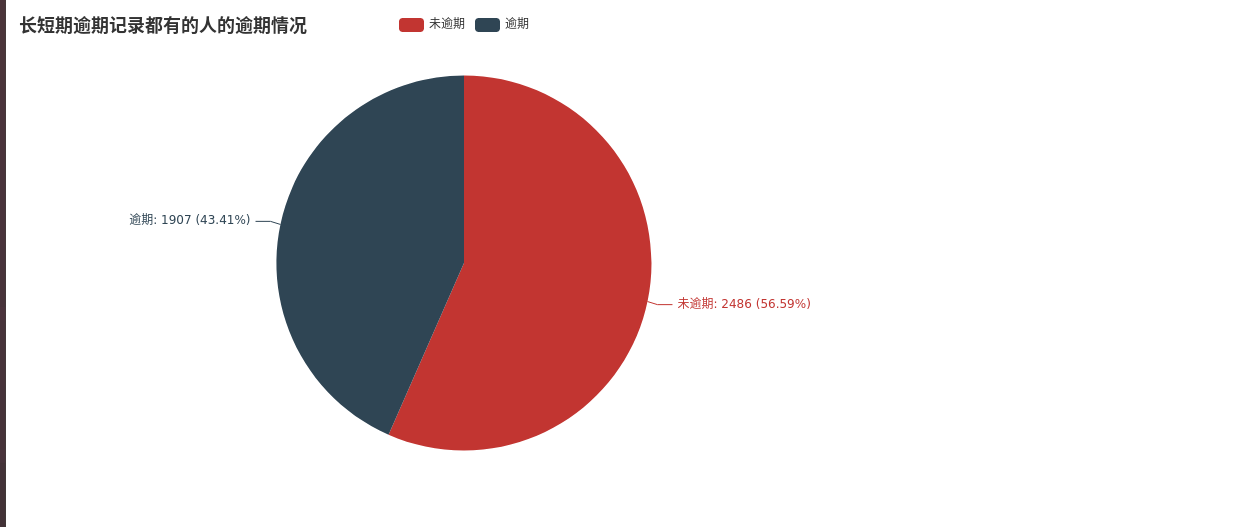




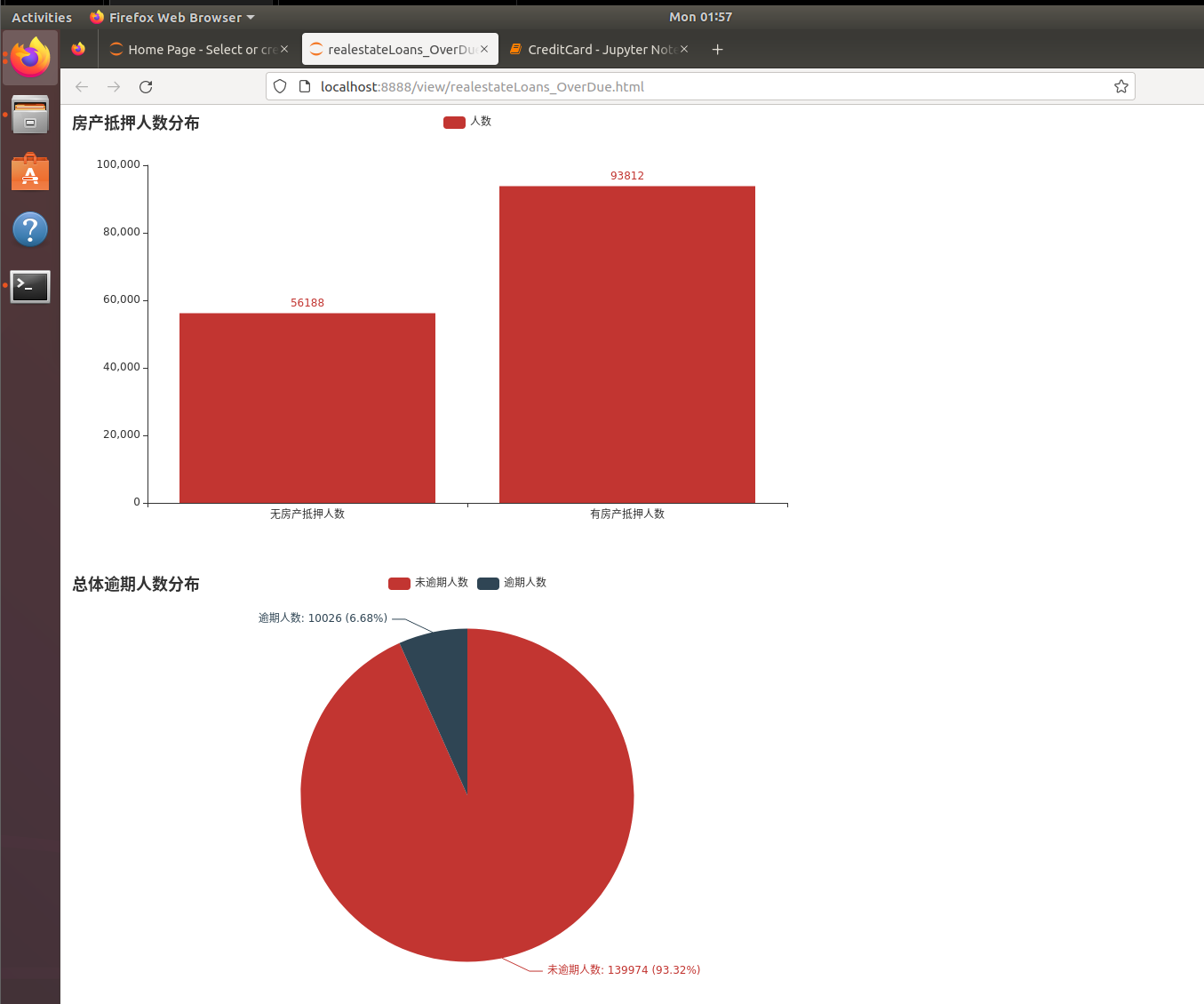
5.3逾期记录展示：







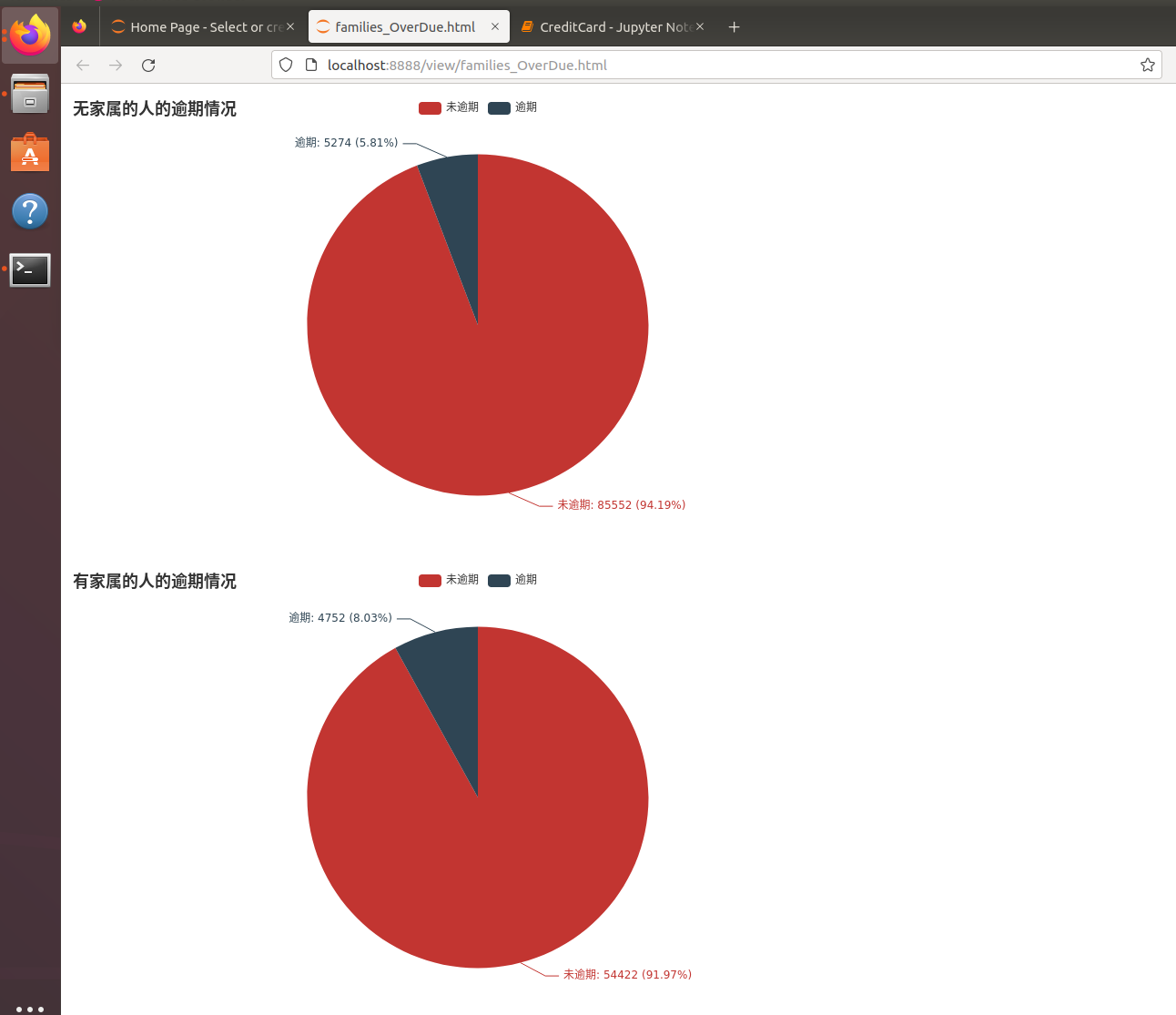
5.4房地产抵押数量展示



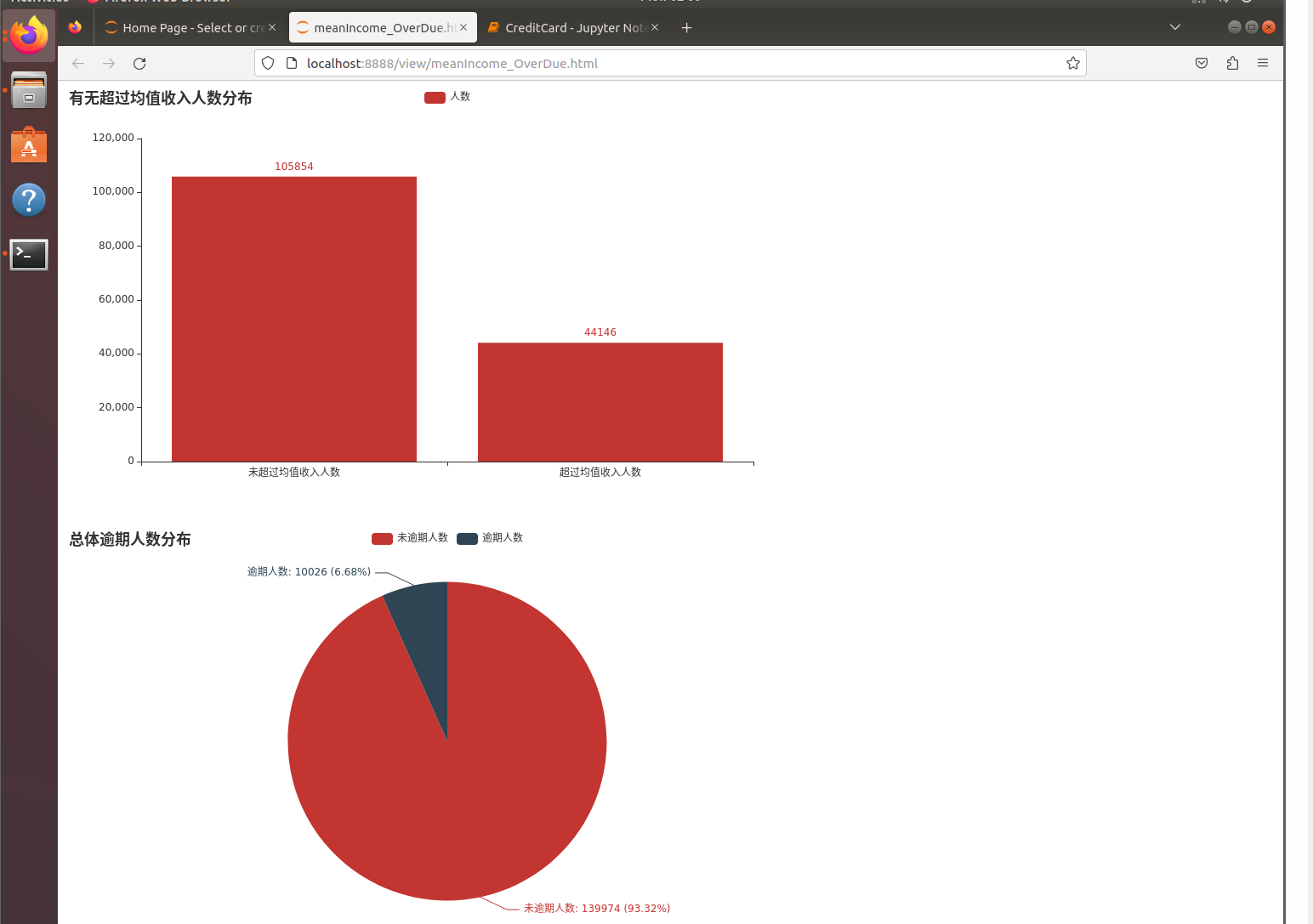


5.5家属人数展示





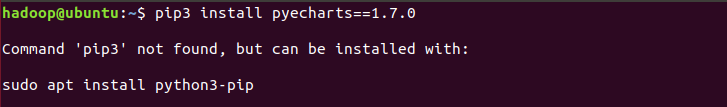
5.6月收入展示



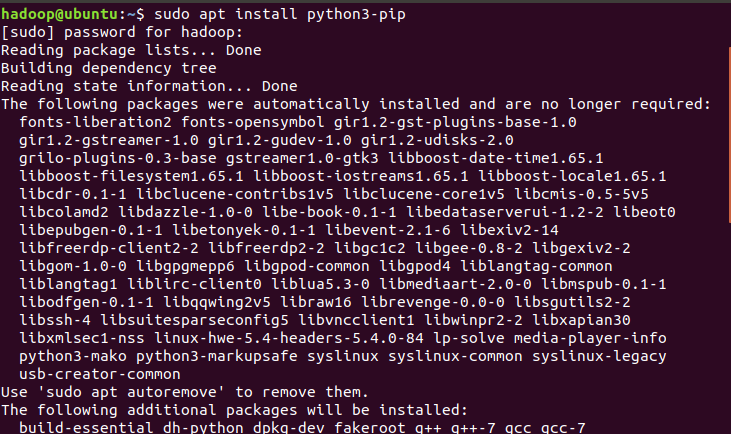


# 二、易错点总结

## （一）无法执行pyecharts安装命令

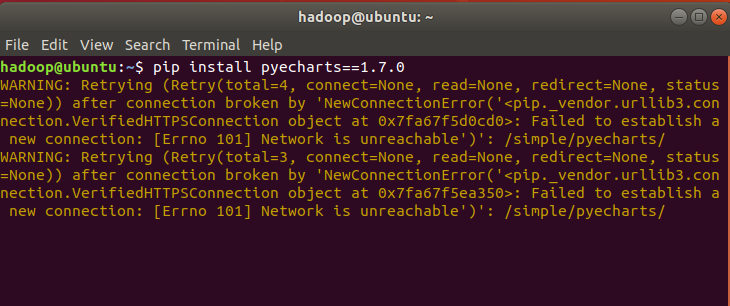


**解决方案：**根据下列提示安装python3-pip，即可执行pip3安装命令





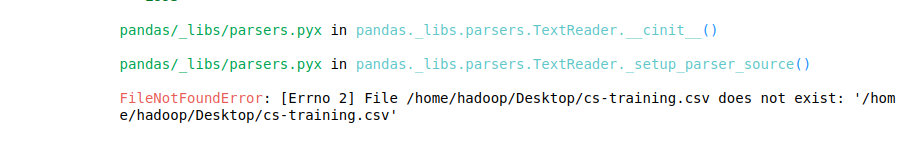
## （二）pyecharts安装无响应

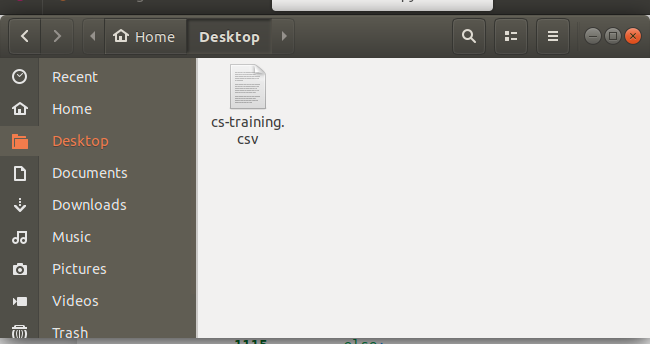


**解决方案：**报错原因是spark没有定义，对spark的定义在数据预处理的代码中可以找到，复制到数据分析中的attribute\_score函数中即可。

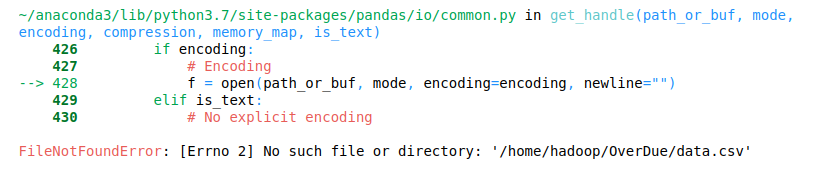


## （三）无法读取原始数据进行预处理

**解决方案：**该报错为路径错误，将代码中的路径改为cs-training.csv中的路径即可（或者也可以把csv文件移动到Desktop目录下，然后再运行代码即可）

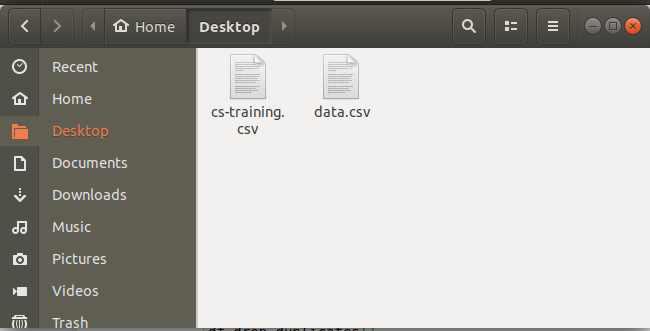


## （四）处理后数据无法保存至本地

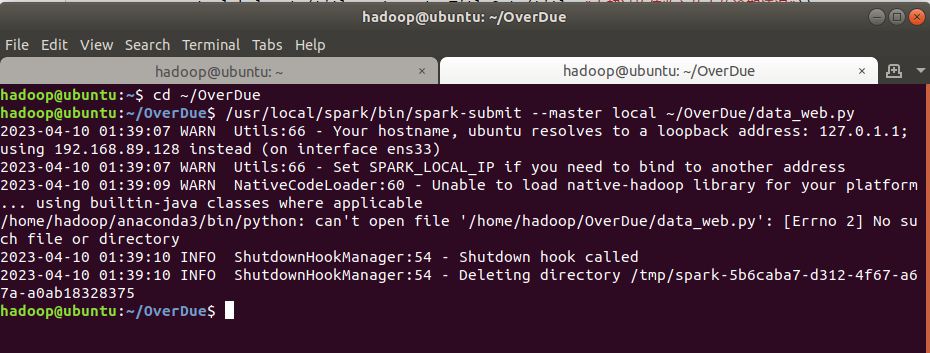


**解决方案：**该报错为文件不存在错误，对应“保存至本地”的那部分代码，我们讲OverDue修改为Desktop（或是任意一个你的虚拟机中存在的目录），代码就可以正常运行，并在对应目录下生成预处理后的数据data.csv。（或者在Home目录下新建OverDue文件夹，即可直接运行指南中的代码，不用进行修改，也不用担心后续的代码修改问题）





## （五）无法将代码文件上传至spark-submit中



**解决方案：**该由于我们没有用到指南中提到的data\_web.py文件，而是直接将代码内容在Jupyter Notebook中运行，所以运行结果直接储存在了Jupyter Notebook中，无需再和指南一样进行上传。

# 三、学习小记

## （一）实验心得

这次的实验环境相较于之前的实验几乎没有改动（感谢老师的仁慈），所以感觉做起来会比较顺手，但正是因为这种想当然的心态，这次在数据处理的过程中也出了很多滑稽的错误。

首先就是路径问题，就比如说用于存放原始数据和清洗后数据的OverDue文件夹是要自己创的，cs-training.csv文件的路径也是在指南代码中被指定好的，而我就是比较懒惰，没好好看代码，等到报错了才一个个去改，也算是为报错集提供了好笑的素材了。

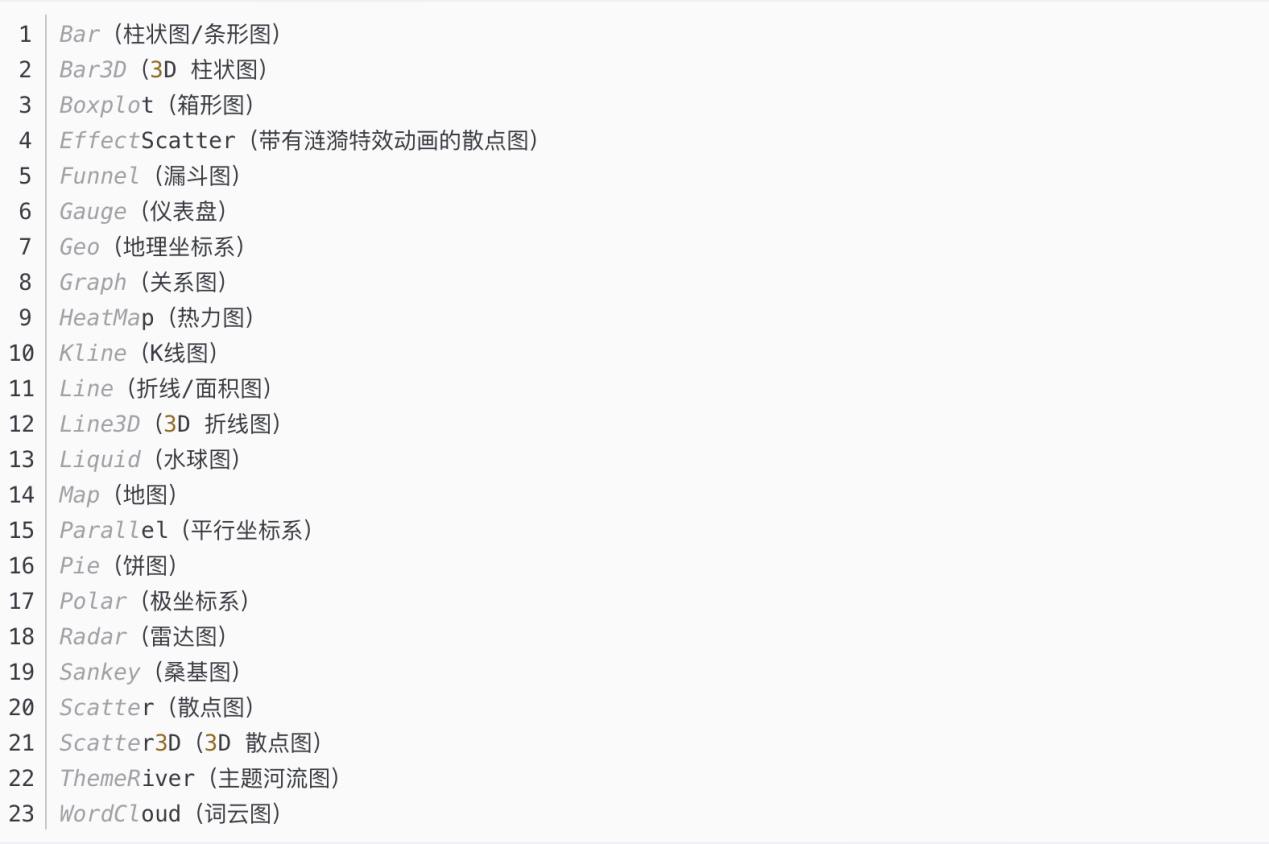
其次是浏览可视化结果的过程中，由于林子雨老师直接使用了代码文件（data\_web.py）进行数据可视化，所以才需要将代码文件提交到spark-submit中生成5个.html的结果文件。而我则是直接在Jupyter Notebook中运行代码，因此根本不需要执行指南最后的shell命令！但我就一直没动那一下脑子，只想依葫芦画瓢，导致报错了都不知道要怎么处理，还是在同学的提醒下才发现原来结果文件就在Notebook里。

总之，实验难度是不大，但也切切实实地给我提了个醒：还是别偷懒，运行之前动动脑子，先把代码和shell命令看一遍，省得报错的时候莫名烦躁。

## （二）知识点笔记

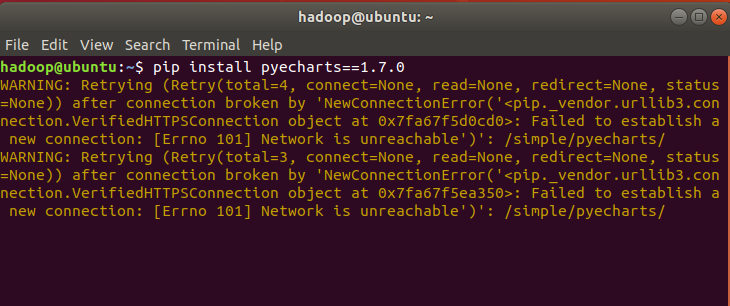
2.1Pyecharts

Pyecharts是一款将python与echarts结合的强大的数据可视化工具，使用pyecharts可以生成独立的网页，也可以在flask、Django中集成使用，包含以下图表：



2.2pip、pip3等命令超时的解决办法

pip请求超时的主要原因是数据源有问题，因为pip默认是使用python官方的源，而国外的官方源常由于网关原因被墙，导致不可用。



遇到超时问题，我们可以尝试如下方法：

1. 更换网络，比如使用自己的热点（笔者此次pyecharts安装超时的问题就是通过使用自己的热点解决的）。
2. 修改源（临时性的），可以在使用pip的时候在后面加上-i参数，指定pip源

pip install scrap -i [https://pypi.tuna.tsinghua.edu.cn/simple](https://pypi.tuna.tsinghua.edu.cn/simple" \t "/Users/yangchaoran/Documents\\x/_blank)

pip install numpy -i [http://pypi.douban.com/simple](http://pypi.douban.com/simple" \t "/Users/yangchaoran/Documents\\x/_blank)